『岡山商大論叢』（岡山商科大学）
第55巻第２号　2019年12月
Journal of OKAYAMA SHOKA UNIVERSITY
Vol.55　No.２　December 2019

《論　　説》

局所的な分布を用いた個票データのリスク評価

佐　井　至　道

Risk Assessment of Microdata Using the Distributions in the Local Regions

Shido Sai

１　はじめに

個票データを公開する場合には，直接識別子と呼ばれる氏名や住所などは事前に削除するのが一般的であるが，その情報のみでは直ちに個体の特定には至らないものの，組み合わせることによって特定される可能性のある準識別子と呼ばれる情報についても，秘匿措置を施すことが多い．

秘匿措置はカテゴリの併合やトップコーディングのような非攪乱的方法と，ノイズの挿入やスワッピングのような攪乱的方法に大別され，これまで国内では前者を用いることが多かった．非攪乱的方法については Willenborg and de Waal（1996, 2001）を，非攪乱的方法で秘匿された個票データのリスク評価で主流となっている寸法指標の利用については佐井（2003, 2013）を参照していただきたい．

しかし，個票データに含まれる情報を残しながら十分安全なデータを作成するためには非攪乱的方法のみでは限界があり，国内でも攪乱的方法を用いることが増えてきている．

ノイズの挿入のような攪乱的方法を施した個票データのリスク評価として，いくつかの方法が提案されているが，主要な方法の１つが距離計測型リンケージと呼ばれるものである．これは，秘匿後の個体のキー変数の値から秘匿前のすべての個体のキー変数の値までの距離を計算して，元の個体と対応づけが可能かどうかを判断する方法である．距離計測型リンケージについては伊藤（2010）を，それ以外のリスク評価方法については伊藤他（2014）を参照していただきたい．

佐井（2015a，2015b，2016）では，キー変数に対してノイズを挿入した場合に，距離計測型リンケージによって元の個体にリンクされる確率の期待値を真のリンク確率の期待値と呼び，リスク評価の指標とした．また，個票データが標本調査で得られたことを想定し，ノイズを挿入したある個体が，標本のみならず母集団でも真のリンクとなる確率について検討を行った．その際，キー変数，ノイズ変数が，主に量的な連続型分布に従う場合を想定した．

佐井（2017b，2017c，2018a，2018b）では，キー変数，ノイズ変数として，連続型と離散型の量的変数を設定して，変数の型によるリスクの違いについて検討を行うとともに，離散型の場合について詳細な検討を行った．

これらの貢献は，ノイズを挿入した個票データのリスク評価において母集団まで考慮に入れた点であるが，キー変数が従う分布を設定する際，一様分布や正規分布のように定義域全体を１つの分布で表現することを考えていたため，適用できるデータが限られていた．本稿では局所的な分布に注目し，多重寸法指標を導入することにより，真のリンク確率の新たな推定方法を提案する．

まず第２章では，ノイズの挿入方法と距離計測型リンケージに基づくリスク評価方法について簡単に説明を行う．キー変数へのノイズの挿入方法

と母集団における個体のリンクについて２.１節で説明し，本稿でリスクの指標として用いる真のリンク確率の期待値を２.２節で説明する．次に第３章では, 佐井（2016, 2018a）などの検討結果について述べるとともに, その貢献と問題点についても説明する.

第４章では，本稿で提案する真のリンク確率の期待値の推定において重要な周辺セルの定義を述べ，対象となる中央セルのサイズと周辺セルのサイズを組み合わせた多重寸法指標について説明を行う.

第５章では，リスク評価方法の手順を３つに分けて説明する．５.１節ではStep １として多重標本寸法指標のカウント方法を述べる．次に５.２節ではStep ２として，多重標本寸法指標を基にした多重母集団寸法指標の推定について述べる．５.３節ではStep ３として，推定された多重母集団寸法指標を用いた真のリンク確率の期待値の推定方法について説明する.

第６章では，人工的な２種類の数値例に対して提案した方法を含むいくつかの方法を用いて推定を行い，提案した方法の有効性について検討を行う.

２　個体のリンクとリスクの指標

２.１　ノイズの挿入と個体のリンク

本稿を通じて，個票データは大きさnの標本から作られたものとする．また，大きさNの母集団においても同じ変数から構成される仮想的な個票データを考える．この２つを単に標本，母集団と呼ぶことがある．ただし，$n = N$とすることによって全数調査のデータにも適用できる.

この節では，標本と母集団における個体のキー変数，標本の個体のキー変数に挿入するノイズ，ノイズが挿入された個体の元の個体へのリンクについて簡単に説明する．詳細については佐井（2016，2017a，2018a）を参照していただきたい.

個票データを構成する変数のうちキー変数の数をKとする．すべてのキー変数を離散型の量的変数とするが，カテゴリに順序のある質的変数に

ついても同様の議論が可能である．個票データでは変数の桁数があらかじめ決まっている場合が多いため，連続型の量的変数については考えないことにする．一方，カテゴリに順序のない質的変数については別途考える必要がある．

また，第三者が母集団に含まれるすべての個体のキー変数の値について完全な情報を持っており，標本のすべての個体のキー変数にノイズを挿入する場合を考える．一部のキー変数の値について情報を持っている場合や，一部のキー変数のみにノイズを挿入する場合については，佐井（2018a）を参照していただきたい．

標本の i 番目 $(i = 1, 2, \ldots, n)$ の個体のキー変数ベクトルと，そのベクトルに挿入するノイズ変数ベクトルをそれぞれ

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,K})' , \tag{1}$$

$$\mathbf{e}_i = (e_{i,1}, e_{i,2}, \ldots, e_{i,K})' \tag{2}$$

として，キー変数の値を $\mathbf{x}_i + \mathbf{e}_i$ として個票データを公開することにする．また，母集団の i 番目 $(i = 1, 2, \ldots, N)$ の個体のキー変数ベクトルを

$$\mathbf{a}_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,K})' \tag{3}$$

とする．なお，標本の i 番目の個体が母集団では i' 番目の個体であるとする．

2つの個体のキー変数ベクトル間の距離を $\mathrm{d}(\cdot, \cdot)$ と表す．一方，あるいは両方のキー変数ベクトルにノイズ変数ベクトルが挿入されている場合も同様に定義する．なお本稿では，すべてのキー変数があらかじめ規準化されているものとして，距離としてはユークリッド距離を用いる．このとき

$$\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{a}_{i''}) \leq \mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i) \tag{4}$$

となる $\mathbf{a}_{i''}$ $(i'' \neq i')$ が少なくとも 1 個存在すれば，キー変数ベクトルにノイズ変数ベクトルを挿入した標本の i 番目の個体について，母集団内で間違ったリンクが発生し，そのような $\mathbf{a}_{i''}$ が存在しなければ，母集団内で真のリンクが発生したと考える．これが佐井（2016）で考えた真のリンクであり，距離計測型リンケージを母集団に拡張したものとなっている．通常

の距離計測型リンケージについては，伊藤（2010），伊藤他（2014）を参照していただきたい．

$\mathbf{x}_i$，$\mathbf{e}_i$ が得られたとき，$\mathbf{a}_{i''}$ がその領域に入ると間違ったリンクとなる範囲を領域Dと呼ぶ．キー変数が連続型の量的変数の場合には，領域Dは，$K=2$ の場合には円，$K=3$ の場合には球，$K\geq 4$ の場合には超球になるが，離散型の量的変数の場合には格子点の集合になる．

図1に，$K=2$ で $x_{i,k}$ が値 $1,2,\ldots,10$ をとりうる場合のイメージを示す．図中の100個のセルは2つのキー変数の値の組み合わせの格子点を表している．以後の図でも格子点の代わりにセルを用いる．

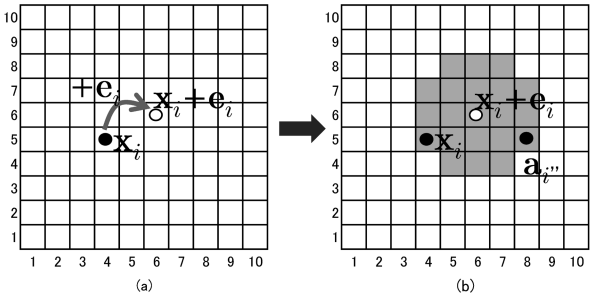

図1　領域Dと個体のリンクのイメージ（$K=2$）

（a）ではキー変数ベクトル $\mathbf{x}_i=(4,5)$ にノイズ変数ベクトル $\mathbf{e}_i=(2,1)$ を加えている．（b）の網掛けされた部分は領域Dを表し，母集団のキー変数ベクトル $\mathbf{a}_{i''}=(8,5)$ が領域D内に存在している．この $\mathbf{a}_{i''}$ は（4）式の等号を成立させているが，このような場合でも，第三者はキー変数の値が $\mathbf{x}_i+\mathbf{e}_i$ として公開された個体が，元の個体であることに確信を持てなくなるため，間違ったリンクが発生したと考える．

なお，キー変数ベクトルが有限領域内しか動けない場合，$\mathbf{x}_i$ がその領域の境界近くにあるとノイズ変数ベクトルを加えた点が領域外に出る可能性がある．また仮にノイズ変数ベクトルを加えた点が領域内の場合でも，領域Dの一部が領域外にはみ出す可能性もある．本稿ではノイズが小さい

ことを想定して，そのようなケースについては無視をして議論する．

2.2 リスクの指標

前節で個体のリンクについて考えたが，リスクの指標としては佐井（2018a）などと同様に，個体のリンクを基礎とした真のリンク確率の期待値を採用する．ただし記法を多少変更している．

標本の i 番目の個体のキー変数ベクトル $\mathbf{x}_i$ が与えられたとする．そのキー変数ベクトルにノイズ変数ベクトル $\mathbf{e}_i$ を挿入したとき，キー変数ベクトル $\mathbf{a}_{i''}$ を持つ母集団の特定の個体が間違ったリンクの 1 つとなる確率，すなわち領域Dに特定の 1 個の個体が入る確率を $p_f(\mathbf{x}_i, \mathbf{a}_{i''})$ と表す．ここでの確率変数は $\mathbf{e}_i$ のみである．

このとき $\mathbf{x}_i$ に対応する母集団の $\mathbf{a}_{i'}$ 以外の $N-1$ 個のキー変数ベクトルが間違ったリンクとならない確率，すなわち真のリンクとなる確率は

$$P_t(\mathbf{x}_i) = \prod_{i''\,(i''\neq i')} \{1 - p_f(\mathbf{x}_i, \mathbf{a}_{i''})\} \tag{5}$$

と書くことができる．簡単のため P_t と表すことがある．この値が大きいほどリスクは大きい．

さらに $\mathbf{x}_i$ について

$$\mathrm{E}\big[P_t(\mathbf{x}_i)\big] = \frac{1}{n}\sum_{i=1}^{n} P_t(\mathbf{x}_i) \tag{6}$$

と期待値をとり，標本の各個体の平均的な真のリンク確率をリスクの指標とする．簡単のため $\mathrm{E}(P_t)$ と表し，「真のリンク確率の期待値」と呼ぶことがある．

なお，(5)，(6) 式では母集団の N 個の個体が与えられた下で定義を行っているが，母集団のキー変数ベクトルに分布を入れて議論することがある．

3　キー変数ベクトルの定義域全体についての分布と局所的な分布

前章で述べた変数のうち，標本のキー変数ベクトル $\mathbf{x}_i$ は観測されており，ノイズ変数ベクトル $\mathbf{e}_i$ はデータを公開する側でコントロールすることができるものの，母集団のキー変数ベクトル $\mathbf{a}_{i''}$ については未知である．したがって，真のリンク確率の期待値を推定するためには，未知である母集団のキー変数ベクトルの分布などを考える必要がある．

佐井（2016）では，キー変数ベクトルが従う分布として多変量連続型一様分布や多変量正規分布など，主に連続型の量的変数の場合を想定して，ノイズの大きさや母集団の大きさなどを変えて真のリンク確率の期待値の比較を行った．解析的に求められない場合にはシミュレーション実験を用いた．

ノイズが大きいほど真のリンク確率の期待値が小さく，すなわちリスクが小さくなり，母集団の大きさ N が大きいほどリスクが小さくなるなど，直感と整合性のある結果が得られた．また，キー変数の個数 K が増えるとリスクが大きくなる結果も得られたが，その増加の速度は極めて速く，経験から得られる直感とは乖離していた．また，分布によるリスクの差は小さいものの，正規分布のように裾を引く分布にキー変数ベクトルが従うと仮定した場合には，連続型一様分布に従うと仮定した場合と比較して，リスクが若干大きくなることも示された．

佐井（2018a）では，キー変数ベクトルが連続型と離散型の量的変数の場合について，真のリンク確率の期待値の比較を行った．佐井（2016）の結果で分布による違いがそれほど大きくなかったため，連続型一様分布，離散型一様分布などに絞って検討を行ったが，解析的に求められない場合には区分求積法を用いた．

その結果，キー変数ベクトルが離散型で，とりうる値が少なくなるほど，連続型の場合と比較してリスクが極端に小さくなることが分かった．ただし，キー変数の個数 K が増えるとリスクが大きくなる傾向は離散型の分布でも同じで，リスクが小さくなるためには，ノイズが十分大きいか，母

集団が十分大きいことが必要になる．

上記の検討の多くでは，キー変数ベクトルの各要素が互いに独立であることを想定していたが，実際のデータでは，キー変数の間に様々な種類の関係性があることが想定される．図２は図１と同様に，$K=2$ で $x_{i,k}$ が値 $1,2,\ldots,10$ をとりうる場合について，標本のキー変数ベクトルの３つの例を示す．図中の点は標本の各個体の値を表す．

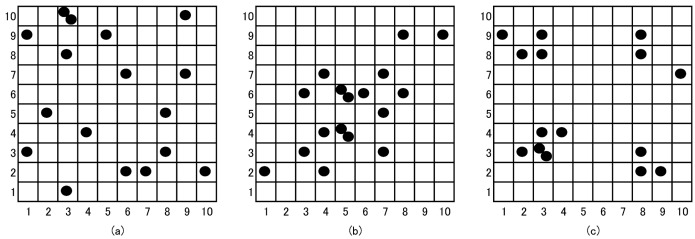

図２　キー変数ベクトルの３つの例（$K=2$）

（a）のような分布の場合には，離散型一様分布を仮定することで，真のリンク確率の期待値を比較的高い精度で推定することが可能と判断される．実際，佐井（2018c）では，アメリカのセンサスの10％抽出個票データ（U. S. Census Bureau（2014））において，年齢，出生四半期（１～３月生まれなど）をキー変数と見なしてノイズを挿入したとき，これらの変数に独立な離散型一様分布を仮定することによって，十分高い精度の推定を行うことが可能だった．

（b）では，２つのキー変数に相関構造があり，例えば多変量正規分布の当てはめが考えられる．佐井（2016）では，キー変数ベクトル，ノイズ変数ベクトルともに同じ相関を持つ２次元正規分布に従う場合の検討を行ったが，結論としては，第三者が相関構造について十分な情報を持っている場合は，リスクはキー変数間に相関がない場合と変わらず，第三者が全く情報を持っていない場合にはリスクは小さくなり，$K=1$ の場合のリスクに近くなることが示された．検討したのは正規分布のみであるが，同じ

ような相関構造を持つ他の分布についても同様の性質が得られることが推測される.

しかし(c)のような分布の場合, これまで紹介したような理論分布をキー変数ベクトルに当てはめるのは難しい. 実際のデータでは, 単純な相関構造を持つだけでなく, 集落のように個体が集まる場所があったり, 逆に個体がほとんど存在しない場所があったりすることが多い. 13歳の既婚者のように, 理論的にそのセルに個体が存在することがない構造的ゼロセルと呼ばれるセルが固まっている部分も存在しうる.

次章では, このような場合に対応するため, キー変数ベクトルの定義域全体について単一の分布を仮定するのではなく, 対象となる個体の周辺のセルにおける局所的な分布に着目した, 真のリンク確率の期待値の新たな推定方法を提案する.

4　周辺セルと多重寸法指標

この章では, 図2の（c）に示したキー変数ベクトルのうち, $\mathbf{x}_i = (8, 8)$ を例にしながら, 局所的な分布を用いたリスク評価方法について説明する. なお簡単のため, 各ノイズ変数 $e_{i,k}$ は ± 1 の値を確率1/2で独立にとる場合を例とする. すなわち, $\mathbf{e}_i = (-1, -1), (-1, 1), (1, -1), (1, 1)$ を確率1/4ずつでとることになる.

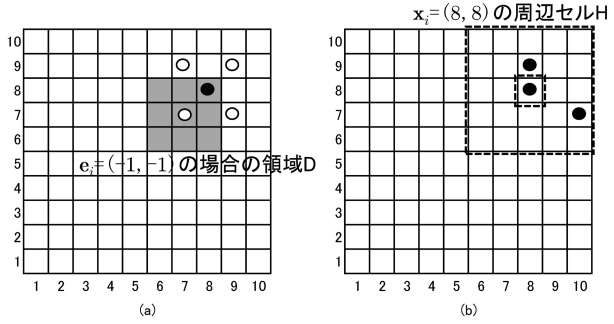

図3　(a) $\mathbf{e}_i$ = (−1, −1) の場合の領域D (b) $\mathbf{x}_i$ = (8, 8) の周辺セルH

図３の（a）では$\mathbf{x}_i = (8,8)$のみを黒い点で，ノイズ変数ベクトルを挿入した後の$\mathbf{x}_i + \mathbf{e}_i$として可能性のある値を白抜きの点で示している．$\mathbf{e}_i = (-1,-1)$の場合，領域Dは図の網掛けされた部分となる．$\mathbf{e}_i = (-1,1),(1,-1),(1,1)$の場合についての領域Dも合わせて，４つの領域Dの和集合を周辺セルHと呼ぶことにする．（b）では，周辺セルHを大きく点線で囲んでいるが，キー変数ベクトルの値が含まれている小さな点線内の中央のセル$(8,8)$は周辺セルには含まれないものとする．周辺セルとの対比で，このセルを「対象とする中央セル」，あるいは単に「中央セル」と呼ぶことがある．なお，ここでは領域D，周辺セルHとも正方形となっているが，ノイズ変数ベクトルの設定により形状は様々である．

セルに含まれる個体数をサイズと呼ぶ．$(8,8)$の中央セルには１個の個体が含まれサイズ１で，その周辺セルには２個の個体が含まれサイズ２となっており，そのサイズの組を$(1,2)$と書くとき，標本におけるそのような組の数を$s_{(1,2)}$と表す．

一般に標本において，対象となる中央セルのサイズが$l\,(l = 0,1,\ldots,n)$，その周辺セルHのサイズが$h\ (h = 0,1,\ldots,n)$の組(l,h)の数を$s_{(l,h)}$と表し，多重標本寸法指標と呼ぶ．母集団でも同様に，中央セルのサイズが$l\,(l = 0,1,\ldots,N)$，その周辺セルHのサイズが$h\ (h = 0,1,\ldots,N)$の組(l,h)の数を$S_{(l,h)}$と表し，多重母集団寸法指標と呼ぶ．

正確には次のように定義する．総セル数がJで，対象となる中央セルがそのj番目$(j = 1,2,\ldots,J)$として，そのセルに含まれる個体数を標本ではf_j，母集団ではF_jとする．また，その周辺セルHに含まれるセル番号の集合を$H(j)$とする．このとき多重標本寸法指標と多重母集団寸法指標は，

$$s_{(l,h)} = \sum_{j=1}^{J} \mathbb{I}(f_j = l,\ \sum_{j' \in H(j)} f_{j'} = h), \tag{7}$$

$$S_{(l,h)} = \sum_{j=1}^{J} \mathbb{I}(F_j = l, \sum_{j' \in H(j)} F_{j'} = h) \tag{8}$$

と書ける．ただし$\mathbb{I}(\cdot)$は定義関数で，（ ）の中がすべて正しければ1を，そうでなければ0を表す関数とする．

サイズの組には，中央セルのサイズlが0の場合，その周辺セルHのサイズhが0の場合，その両方が成り立つ場合も含まれる．

対象となる中央セルだけに着目したとき，サイズがlとなるセル数を，標本ではs_lと表して標本寸法指標と呼び，母集団ではS_lと表して母集団寸法指標と呼ぶ．秘匿措置が施されていない個票データや非攪乱的な秘匿方法が施された個票データでは，標本寸法指標を基に母集団寸法指標を推定するのが，リスク評価方法の主流となっている．多重寸法指標は寸法指標の拡張になっており，次の式が成り立つ．

$$s_l = \sum_{h=0}^{n} s_{(l,h)}, \tag{9}$$

$$S_l = \sum_{h=0}^{N} S_{(l,h)}. \tag{10}$$

同様に，多重標本寸法指標を基に多重母集団寸法指標を推定できれば，対象となる個体が入る中央セルの近くにおける母集団での個体の集散状況が推測できるため，真のリンク確率の期待値を推定することが可能となる．逆に真のリンク確率の期待値を推定するためには，キー変数ベクトルの定義域全体について表現できる単一の分布でなく，局所的な分布を考えるだけで十分というのが，この推定方法の提案の基礎にある．

なお，前述したように，領域Dがキー変数ベクトルの定義域からはみ出る可能性があるため，周辺セルHもまたはみ出る可能性がある．この点に

ついては無視をして説明する.

5　局所的な分布を用いたリスク評価方法

5.1　多重標本寸法指標の計算

提案するリスク評価方法では，次の3つのステップで真のリンク確率の期待値を推定することになる.

Step 1　多重標本寸法指標$s_{(l,h)}$をカウントする
Step 2　$s_{(l,h)}$を基に多重母集団寸法指標の推定値$\hat{S}_{(l,h)}$を求める
Step 3　$\hat{S}_{(l,h)}$を用いて真のリンク確率の期待値$\mathrm{E}(P_t)$を推定する

この節ではStep 1について考えるが，多重標本寸法指標のカウントについて検討する前に，標本寸法指標s_lをカウントする場合について簡単に説明する.

理論的にはすべてのセルについて，そのセルに含まれる個体数をカウントすれば良いものの，実際のデータではキー変数の個数Kが大きいことが多いため，総セル数Jも極めて大きくなり，この方法では計算時間的に標本寸法指標のカウントが困難になる場合が多い．そのため，個体が含まれるセルのみを対象にして，標本寸法指標をカウントしなければならない．実際には，その個体のキー変数ベクトルと同じものが，標本に含まれるすべての個体の中で何個あるか数える方法がとられることが多い．なおs_0については，総セル数Jからサイズ1以上のセル数を引くことによって求めることができるが，母集団寸法指標の推定に用いるモデルによっては，s_0の値が不要なものもある.

これに対して，後に用いる多重母集団寸法指標の推定法では，$l=0$かつ$h=0$の多重標本寸法指標$s_{(l,h)}$をカウントする必要はないものの，$l=0$かつ$h>0$，あるいは$l>0$かつ$h=0$の多重標本寸法指標についてはカウントしなければならないため工夫が必要となる．例えば次のような

方法が考えられる．

ここでも，各ノイズ変数$e_{i,k}$は±１の値をとるものとして，周辺セルは図３の（b）で考えたものと同じ大きさの場合を例に考える．

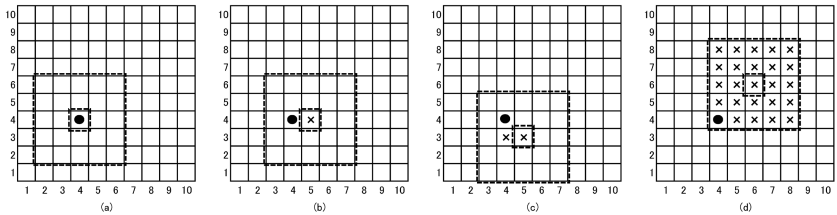

図４　多重標本寸法指標のカウントの工夫の例

まず対象となる中央セルのうち，図４の（a）のようにサイズ１以上となるセルのみについて多重標本寸法指標をカウントする．このようなものは高々n個である．

周辺セルHに含まれるセル数を$\#(\mathrm{H})$として，それらのセルに中央セルとの位置関係で１から$\#(\mathrm{H})$の順番を決める．次に，（b）のように中央セルのサイズが０で，周辺セルの１番目のセルのサイズが１以上のものをカウントする．×印はサイズ０を意味する．このようなものも高々n個である．（b）では中央セルの左隣のセルを周辺セルの１番目としている．さらに，（c）のように中央セルと周辺セルの１番目のサイズがともに０で，周辺セルの２番目のセルのサイズが１以上のものをカウントする．（c）では中央セルの左上のセルを周辺セルの２番目としている．

以後同様にカウントしていき，最後は，（d）のように周辺セルの$\#(\mathrm{H})$番目のセルのみのサイズが１以上で，中央セルを含めてそれ以外のすべてのセルのサイズが０のものをカウントすることになる．（d）では周辺セルの左下隅のセルを周辺セルの$\#(\mathrm{H})$番目としている．

このような方法を用いると，カウントする回数は，高々$(\#(\mathrm{H})+1)\cdot n$回となる．ノイズが大きい場合やキー変数の個数が多い場合には$\#(\mathrm{H})$も大きくなるが，総セル数Jと比較するとカウントの回数を大幅に減らすこと

が可能である.

5.2 多重母集団寸法指標の推定

多重標本寸法指標$s_{(l,h)}$を基に多重母集団寸法指標$S_{(l,h)}$を推定する方法としては，モデルを用いるパラメトリックな方法とノンパラメトリックな方法が提案されている.

渋谷，佐井（2007），Sibuya and Sai（2008）は，母集団寸法指標の推定に用いられているピットマンモデルを多重寸法指標に拡張した多重ピットマンモデルを用いて多重母集団寸法指標を推定する方法を提案した．ただし，このモデルは$s_l = \sum_h s_{(l,h)}$，$s_h = \sum_l s_{(l,h)}$の２つの寸法指標が同じパラメータの同じモデルで表され，lとhについて対称なものであった．例えば，同じ人数の男女について，あるセルに含まれる男性の人数をl，女性の人数をhとする場合や，同じ集団について，ある時点にあるセルに含まれる個体数をl，別の時点に同じセルに含まれる個体数をhとする場合については適用可能であるが，ここで扱うような非対称な場合については適用が難しい．なおピットマンモデルについてはPitman（1995）を参照していただきたい.

そのため本稿では，佐井（2008，2009）で提案したノンパラメトリック最尤推定法を用いることにする．ここでは理論の詳細は省き，本稿の設定に合わせた推定法の概略を述べることにする．詳細については佐井(2009)を参照していただきたい.

まず，$(l,h) = (0,0)$を除くサイズの組について，多重標本寸法指標を並べたベクトルと多重母集団寸法指標を並べたベクトルを

$$\boldsymbol{s} = (s_{(0,1)}, s_{(0,2)}, \ldots, s_{(1,0)}, s_{(1,1)}, \ldots), \tag{11}$$

$$\boldsymbol{S} = (S_{(0,1)}, S_{(0,2)}, \ldots, S_{(1,0)}, S_{(1,1)}, \ldots) \tag{12}$$

と表す.

大きさNの母集団から大きさnの標本が非復元単純無作為抽出された場合，抽出率を$\lambda = n/N$とするとき，多重標本寸法指標の組$\boldsymbol{s}$が得られ

たときの多重母集団寸法指標の組$\boldsymbol{S}$の尤度関数は

$$\begin{aligned}&\mathrm{L}(\ \boldsymbol{S}\mid\boldsymbol{s}\ )\\ =\ &\frac{1}{\binom{N}{n}\binom{m/\lambda}{m}}\sum_{\mathcal{C}_1}\prod_{(l,h)}\left\{\frac{S_{(l,h)}!}{\prod_{(l',h')}k_{(l,h),(l',h')}!}\prod_{(l',h')}\left\{\binom{l}{l'}\binom{h}{h'}\right\}^{k_{(l,h),(l',h')}}\right\}\end{aligned}\tag{13}$$

と表される．ただし$m=\#(\mathrm{H})\cdot n$は標本における周辺サイズの合計である．また$k_{(l,h),(l',h')}$は，母集団においてサイズの組(l,h)のうち標本においてサイズの組(l',h')となる数を表し，$\mathcal{C}_1$は$\boldsymbol{S}$から$\boldsymbol{s}$が生成されるような$k_{(l,h),(l',h')}$のすべての組み合わせを表す．なお（13）式の積において，母集団におけるサイズの組み合わせ(l,h)は$(0,0)$を含まず，標本におけるサイズの組み合わせ(l',h')は$(0,0)$を含む．

母集団が大きくなるにつれて，（13）式で和をとる組み合わせ$\mathcal{C}_1$の数は膨大となり，計算時間的な困難が生じる．そのため，次のようにポアソン分布の確率関数の積で近似を行う．

まず，母集団の各個体を他の個体と独立に確率λで抽出するベルヌーイ抽出を考える．多重標本寸法指標の組$\boldsymbol{s}$が得られたときに，母集団の大きさをNに限定した多重母集団寸法指標の組$\boldsymbol{S}$の尤度関数は

$$\begin{aligned}&\mathrm{L_B}(\ \boldsymbol{S}\mid\boldsymbol{s}\ )\\ =\ &\sum_{\mathcal{C}_1}\prod_{(l,h)}\Bigg\{\frac{S_{(l,h)}!}{\prod_{(l',h')}k_{(l,h),(l',h')}!}\\ &\qquad\cdot\prod_{(l',h')}\left\{\binom{l}{l'}\lambda^{l'}(1-\lambda)^{l-l'}\cdot\binom{h}{h'}\lambda^{h'}(1-\lambda)^{h-h'}\right\}^{k_{(l,h),(l',h')}}\Bigg\}\\ =\ &\binom{N}{n}\lambda^n(1-\lambda)^{N-n}\cdot\binom{m/\lambda}{m}\lambda^m(1-\lambda)^{m/\lambda-m}\cdot\mathrm{L}(\ \boldsymbol{S}\mid\boldsymbol{s}\ )\end{aligned}\tag{14}$$

と書ける．尤度関数の添え字のBはベルヌーイ抽出を表す．多重母集団寸

法指標によらず，ベルヌーイ抽出の尤度関数は非復元単純無作為抽出の尤度関数の定数倍になるため，母集団の大きさがNの中で（14）式を最大とする$\boldsymbol{S}$を求めればそれが（13）式の最尤推定値となる．

ここで，抽出率λが十分小さい場合には，ベルヌーイ抽出の尤度関数はポアソン分布の確率関数で近似することができ，最終的に，非復元単純無作為抽出の尤度関数は次のように近似することができる．

$$\mathrm{L_{ap}}(\ \boldsymbol{S} \mid \boldsymbol{s}\ ) = \frac{1}{\binom{N}{n}\lambda^n(1-\lambda)^{N-n} \cdot \binom{m/\lambda}{m}\lambda^m(1-\lambda)^{m/\lambda-m}} \cdot \prod_{(l',h')} \frac{e^{-\mu_{(l',h')}}\mu_{(l',h')}^{s_{(l',h')}}}{s_{(l',h')}!}. \tag{15}$$

ただし，

$$\mu_{(l',h')} = \sum_{(l,h)(\geq(l',h'))} \left\{ S_{(l,h)} \cdot \binom{l}{l'}\lambda^{l'}(1-\lambda)^{l-l'} \cdot \binom{h}{h'}\lambda^{h'}(1-\lambda)^{h-h'} \right\} \tag{16}$$

である．なお，（15）式の積において，標本におけるサイズの組み合わせ(l',h')は$(0,0)$を含まない．

（15）式を$S_{(l,h)}$で偏微分して0とすることによって，多重母集団寸法指標の近似的な最尤推定値を求めることは可能だが，推定は非常に不安定で負の値が含まれることが多い．そのため，後で述べるように，多重母集団寸法指標の推定値に制約を入れ，探索法によって近似的な解を求めることにする．

5.3　真のリンク確率の期待値の推定

この節では，推定された多重母集団寸法指標を基にした真のリンク確率の期待値$\mathrm{E}(P_t)$の推定について考える．

情報を寸法指標まで縮約すると，同じセルに含まれる母集団と標本の個体数の対応づけができなくなるため，ここでは多重標本寸法指標は多重母

集団寸法指標の推定のみに用い，真のリンク確率の期待値の推定は，推定された多重母集団寸法指標の情報のみに基づいて行う．具体的には，母集団の個体が母集団において真のリンクとなる確率の期待値によって，標本の個体が真のリンクとなる確率の期待値を推定する．

多重母集団寸法指標の推定値$\hat{S}_{(l,h)}$が得られたとき，母集団で対象となる中央セルのサイズが1，周辺セルのサイズがh，となる中央セルと周辺セルの組の1つを考える．中央セルのサイズが2以上であれば，そのセルに含まれる個体にどのようなノイズを挿入しても，あるいはノイズを挿入しなくても，同じセルの別の個体が間違ったリンクとなる．そのため，サイズ2以上のセルに含まれる個体が真のリンクとなることはない．

ここで周辺セルH内で母集団の個体が一様に分布すると仮定する．キー変数ベクトルが$\mathbf{a}_i$である対象となる中央セルの個体が，真のリンクとなる確率は

$$\hat{P}_t(\mathbf{a}_i) = \left(1 - \frac{\#(\mathrm{D}) - 1}{\#(\mathrm{H})}\right)^h \tag{17}$$

と推定できる．ただし$\#(\mathrm{D})$は領域Dのセル数を表すが，挿入されるノイズによって$\#(\mathrm{D})$が変動する場合には，期待値をとることになる．前述したように，対象となる中央セルのサイズが2以上の場合は$\hat{P}_t(\mathbf{a}_i) = 0$である．

標本の個体の真のリンク確率の期待値は，母集団の個体の真のリンク確率の期待値と等しく

$$\mathrm{E}\big[\hat{P}_t(\mathbf{x}_i)\big] = \mathrm{E}\big[\hat{P}_t(\mathbf{a}_i)\big] = \frac{1}{N}\sum_{i=1}^{N} \hat{P}_t(\mathbf{a}_i) \tag{18}$$

と推定することができる．

各ノイズ変数$e_{i,k}$が± 1の値を確率1/2で独立にとる場合について，真のリンク確率の計算手順を例示する．

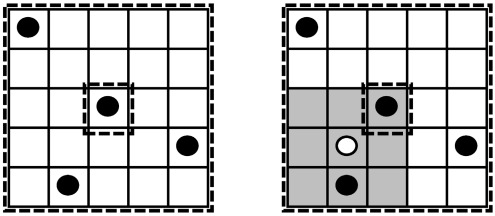

図５　真のリンク確率の期待値の推定のイメージ

図５の（a）に示すように，対象となる中央セルに母集団の個体が１個含まれており，$\#(\mathrm{H}) = 24$のセルからなる周辺セルには母集団の個体が$h = 3$個含まれている．ノイズ変数ベクトルが$\mathbf{e}_i = (-1, -1)$の場合の領域Dは図５の（b）の左下の網掛けされた部分となり，対象となる中央セルを除くとセル数は$\#(\mathrm{D}) - 1 = 8$となる．他の３通りのノイズ変数ベクトルが挿入された場合もセル数は同じである．この図では，$\mathbf{e}_i = (1, 1)$のときには真のリンク，$\mathbf{e}_i = (-1, -1), (-1, 1), (1, -1)$のときには間違ったリンクになるが，周辺セル内で母集団の個体が一様分布するという仮定の下，$h = 3$という情報のみを用いると，この個体が標本としてとられた場合に真のリンクとなる確率は３個の個体が領域Dの外部にある確率として

$$\hat{P}_t(\mathbf{a}_i) = (1 - 8/24)^3 = 0.29630$$

と推定される．

６　数値実験による検討

この章では，人工的な２つのデータ（A），（B）を用いて，提案した方法の有効性について検討する．

（A），（B）ともキー変数の個数は$K = 3$で，各キー変数は１から100までの整数値をとる．また母集団の大きさは$N = 2 \cdot 10^4$，標本の大きさは$n = 10^4$とする．

（A）は各キー変数が独立な離散型一様分布に従うデータとする．一様分布の確率関数を用いてN個の乱数の組を発生させ，これを母集団のキー

変数ベクトルとする．(A) を簡単に一様分布と呼ぶ．

(B) は各キー変数は独立で，確率関数 $f(1), f(2), \ldots, f(5)$ が1/300, 2/300, 3/300, 4/300, 5/300と単調増加で，$f(6), f(7), \ldots, f(10)$ が5/300, 4/300, 3/300, 2/300, 1/300, と単調減少し，以後同様に10毎に周期がある分布に従うデータとする．この確率関数を基に N 個の乱数の組を発生させ，これを母集団のキー変数ベクトルとする．(B) を偏在分布と呼ぶことにする．

２つの母集団から，それぞれ10^4個の個体を非復元単純無作為抽出して，標本とする．次に，標本の各キー変数に ± １のノイズを確率1/2ずつで独立に挿入して，これを公開用の個票データとする．表１に（B）偏在分布の標本について最初の５つの各キー変数の値と，ノイズ変数を挿入した公開用の値を示す．１番目の個体のように，キー変数の値が100の場合には確率１で－１のノイズを入れ，値が０の場合には確率１で１のノイズを入れる．

表１　(B) 偏在分布の標本におけるキー変数と公開用の値

番号	x_1	x_2	x_3	x_1+e_1	x_2+e_2	x_3+e_3
1	47	10	100	46	9	99
2	91	24	5	90	25	4
3	5	39	45	6	40	46
4	66	14	58	65	13	57
5	76	8	75	77	9	74
:	:	:	:	:	:	:

この公開用のデータの各個体について，母集団の中で真のリンクとなるか間違ったリンクとなるかを１個ずつ確認したところ，真のリンクとなった比率は，(A) が0.60330, (B) が0.48900であった．これが推定目標となる．

それぞれの標本について求めた多重標本寸法指標を表２と表３に示す．

表２　（A）一様分布の多重標本寸法指標 $s_{(l,\,h)}$

		h												
		0	1	2	3	4	5	6	7	8	9	10	11	12
	0	300112	356264	213539	86619	26133	6080	1139	159	6	0	0	0	0
l	1	3003	3622	2147	826	246	44	8	2	0	0	0	0	0
	2	11	21	12	5	2	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0	0

表３　（B）偏在分布の多重標本寸法指標 $s_{(l,\,h)}$

		h												
		0	1	2	3	4	5	6	7	8	9	10	11	12
	0	354003	313330	176615	85161	37463	15105	5569	1864	663	205	89	31	7
l	1	2106	2725	2251	1413	737	329	133	63	22	8	3	1	1
	2	11	25	21	16	12	9	5	1	0	1	0	0	0
	3	0	0	0	0	0	1	1	0	0	0	0	0	0

この多重標本寸法指標を基にして，５.２節の方法で多重母集団寸法指標を推定するが，その際，多重母集団寸法指標の推定値に次の制約条件を課す.

(1)　$\hat{S}_{(l,h)} \geq 0 \quad (l = 0, 1, \ldots, N,\ h = 0, 1, \ldots, N)$

(2)　$\hat{S}_{(l,h)} \leq \hat{S}_{(l-1,h)} \quad (l = 1, \ldots, N,\ h = 0, 1, \ldots, N)$

(3a)　$\hat{S}_{(l,h-1)} \leq \hat{S}_{(l-1,h)} \quad (l = 1, \ldots, N,\ h = 1, \ldots, N)$

(3b)　$\hat{S}_{(l,h)} \leq \hat{S}_{(l-1,h-1)} \quad (l = 1, \ldots, N,\ h = 1, \ldots, N)$

(1) は多重母集団寸法指標の推定値が非負であることを，(2) は後に示す多重母集団寸法指標の推定値の表で各列が下に単調減少であることを意味し，(3a)，(3b) は斜め下に単調減少であることを意味する．(1) のみの制約を基本にして，(A) 一様分布については (2) あるいは (2)，(3a)，(3b) を加えた場合も考える.

これらの制約条件を推定に取り入れるために，渋谷 (2005) が提案した滑らかなペナルティー関数

$$\mathrm{Pnlt}(x;\varepsilon) = \varepsilon \log(e^{\frac{x}{\varepsilon}} + 1) \tag{19}$$

を用いる．ただし，εはパラメータである.

（15）式の対数をとった尤度関数に，例えばすべての制約条件を用いる場合には

$$\begin{aligned}\mathrm{F}(\ \boldsymbol{S}\mid\boldsymbol{s}\ ) \ =\ \log \mathrm{L}_{\mathrm{ap}}(\ \boldsymbol{S}\mid\boldsymbol{s}\ ) &- c_1\sum_{(l,h)}\mathrm{Pnlt}(-S_{(l,h)};\ \varepsilon_1)\\ &- c_2\sum_{(l,h)}\mathrm{Pnlt}(-(S_{(l-1,h)}-S_{(l,h)});\ \varepsilon_2)\\ &- c_3\sum_{(l,h)}\mathrm{Pnlt}(-(S_{(l-1,h)}-S_{(l,h-1)});\ \varepsilon_3)\\ &- c_3\sum_{(l,h)}\mathrm{Pnlt}(-(S_{(l-1,h-1)}-S_{(l,h)});\ \varepsilon_3)\end{aligned}\qquad(20)$$

とペナルティー関数を取り込む．ε_1，ε_2，ε_3は滑らかさを表すパラメータであり，c_1，c_2，c_3はそれぞれのペナルティー関数の関数Fへの影響力を表す係数である．

ここでは，これまでの経験から$\varepsilon_1 = 0.001$，$\varepsilon_2 = \varepsilon_3 = 0.1$，$c_1 = 1.0$，$c_2 = c_3 = 1000.0$として，適当な初期値から，ステップごとにFが増加する方向に各$S_{(l,h)}$を独立に移動させる勾配法によって最大値を探索し，得られた近似的な値を多重母集団寸法指標の推定値とする．移動途中は実数値を考え，最大値を求めた時点で丸めて整数値とする．ペナルティー関数と勾配法の詳細については佐井（2009）を参照していただきたい．

表４には（A）一様分布について，推定された多重母集団寸法指標を示す．表４の上の表は制約条件（1）のみを課した場合，下の表は制約条件（1），（2），（3a），（3b）を課した場合である．なお，制約条件として（1），（2）を課した場合の多重母集団寸法指標の推定値は上の表と同じである．また表５には実際の多重母集団寸法指標を示す．

表６には（B）偏在分布について，制約条件（1）のみを課した場合の推定された多重母集団寸法指標を示し，表７には実際の多重母集団寸法指標を示す．

表４　(A) 一様分布の推定された多重母集団寸法指標 $\hat{S}_{(l,\,h)}$

（制約条件 (1) のみの場合）

		h												
		0	1	2	3	4	5	6	7	8	9	10	11	12
	0		206352	247346	205419	124969	82055	11034	17456	1984	0	0	0	0
l	1	1688	3662	6482	2472	3385	1631	0	190	82	0	0	0	0
	2	0	35	85	0	69	16	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0	0

（制約条件 (1)，(2)，(3a)，(3b) の場合）

		h												
		0	1	2	3	4	5	6	7	8	9	10	11	12
	0		206300	247631	204568	126525	80247	12350	16904	2083	2	0	0	0
l	1	1692	3617	6656	2058	3982	1194	27	362	2	1	0	0	0
	2	0	39	85	1	56	26	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0	0

表５　(A) 一様分布の実際の多重母集団寸法指標 $S_{(l,\,h)}$

		h												
		0	1	2	3	4	5	6	7	8	9	10	11	12
	0	91287	216217	253811	201843	121763	59177	24226	8537	2523	660	139	22	1
l	1	1963	4239	5127	4010	2425	1165	441	165	43	10	1	0	0
	2	16	46	50	49	32	7	4	0	0	0	0	0	0
	3	0	0	0	1	0	0	0	0	0	0	0	0	0

表６　(B) 偏在分布の推定された多重母集団寸法指標 $\hat{S}_{(l,\,h)}$

		h																
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	0		230556	204656	128578	138333	0	95161	4654	213	22527	0	0	0	182	1368	0	0	0
l	1	1403	2420	2809	4747	0	4386	2122	0	0	706	538	0	0	0	0	55	5	0
	2	0	26	57	123	0	0	0	120	50	0	0	0	0	0	0	6	0	0
	3	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0

表７　(B) 偏在分布の実際の多重母集団寸法指標 $S_{(l,\,h)}$

		h																	
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	0	160832	229575	200865	146734	97163	60935	37060	21693	12102	6720	3600	1751	818	313	142	57	14	3
l	1	1373	2617	3258	3158	2721	2071	1569	1061	613	389	227	104	56	25	12	5	0	0
	2	10	28	45	49	57	42	49	32	19	10	4	2	4	1	0	0	0	0
	3	0	0	1	0	1	1	2	0	2	1	2	0	1	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

次に 5.3 節の方法で，多重母集団寸法指標の推定値を基に真のリンク確率の期待値 $\hat{P}_t(\mathbf{a}_i)$ を（18）式によって求める．(17) 式において $\#(\mathrm{D}) = 3^3 = 27$，$\#(\mathrm{H}) = 5^3 - 1 = 124$ を用いることになる．なお周辺セル H がキー変数の定義域の外にはみ出した場合には，はみ出した部分には個体が存在しないとみなし，通常の周辺セルと同じ扱いをする．

表８に（A）一様分布の結果を示す．

表８　（A）一様分布の真のリンク確率の期待値の推定値

n	提案した方法	真の指標利用	一様分布利用	実際の比率
10000	0.58619	0.59737	0.58276	0.60330
5000	0.58307	同上	同上	0.60220
2000	0.54247	同上	同上	0.61300

$n = 10000$の行の「提案した方法」に，制約条件（1）のみを課した場合の真のリンク確率の期待値の推定値$\mathrm{E}\big[\hat{P}_t(\mathbf{x}_i)\big] = 0.58619$を示す．また同じ行の「実際の比率」が前述した推定目標である．「真の指標利用」には多重母集団寸法指標として$\hat{S}_{(l,h)}$でなく，実際の値$S_{(l,h)}$を用いた場合の推定結果を，「一様分布利用」には佐井（2018a）などで提案したキー変数が定義域全体で一様分布に従うことを想定した場合の推定結果を示す．

（A）ではキー変数が一様分布に従っているため，本稿で提案した方法も一様分布を利用した方法も推定目標とほとんど変わらず，十分高い精度の推定がなされていると判断できる．制約条件として（1），（2）を課した場合の真のリンク確率の期待値の推定値は$\mathrm{E}\big[\hat{P}_t(\mathbf{x}_i)\big] = 0.58619$，（1），（2），（3a），（3b）を課した場合の推定値は$\mathrm{E}\big[\hat{P}_t(\mathbf{x}_i)\big] = 0.58612$であり，ほとんど差は認められない．制約が弱い場合，多重母集団寸法指標の推定値の表で，正の値の中に０が混在する虫食い状態が発生するが，推定結果への影響は限定的と考えられる．

標本の大きさを$n = 5000$，$n = 2000$とさらに小さくした場合に，同様にして推定された真のリンク確率の期待値と１個ずつ確認した真のリンクの比率をそれぞれの行に示す．$n = 2000$の場合には推定値が推定目標よりも若干小さくなっている．

表９に（B）偏在分布の結果を示す．

表９　(B) 偏在分布の真のリンク確率の期待値の推定値

n	提案した方法	真の指標利用	一様分布利用	実際の比率
10000	0.47103	0.47483	0.58276	0.48900
5000	0.45620	同上	同上	0.49020
2000	0.45068	同上	同上	0.48250

キー変数が従う分布が一様分布から乖離しているため，キー変数が定義域全体で一様分布に従うことを仮定している「一様分布利用」の結果は「実際の比率」と大きく異なっている．一方，本稿で提案した方法では（A）一様分布と同様に十分良い推定が行われていると判断される．

7　おわりに

本稿では，個票データに含まれるレコードのキー変数ベクトルにノイズ変数ベクトルを挿入した場合のリスクとして，距離計測型リンケージを基にした真のリンク確率の期待値を考えた．その推定では，対象となる個体が含まれるセルのサイズとともにその周辺セルのサイズも合わせて考え，その表現として多重寸法指標を用いた．多重標本寸法指標からの多重母集団寸法指標の推定においては，制約条件を課すノンパラメトリック最尤推定法を基にした探索的な方法を用いた．

人工的なデータによる検討結果からは，キー変数ベクトルの定義域全体で単一の分布に従うことを仮定する推定法よりも，優れている場合のあることが読み取れた．一様分布や正規分布などで表現できないキー変数ベクトルは多く，提案した方法が有効なケースは少なくないと考えられる．ただ，検討では，キー変数の個数が３，母集団の大きさが$2 \cdot 10^4$という小さい母集団を用い，抽出率も実際の調査と比較して高かった．官庁統計データへの適用を考えると，さらに大きい母集団についての検討も必要である．また，実際のデータではキー変数ごとにとりうる値の数が異なるのが一般的であり，この点についての検討も必要と考えられる．

多重標本寸法指標からの多重母集団寸法指標の推定においてはノンパラメトリック法を用いたが，母集団寸法指標の推定に用いたこれまでの経験

では，モデルを用いた推定法と比較して推定が不安定であることが多かった．制約条件を推定を安定させるために導入しているが，セルとその周辺セルという組み合わせの多重寸法指標について，実際のデータを扱った経験はわずかである．そのため適切な制約条件を設定するためには，さらなる情報の収集が必要である．

多重母集団寸法指標の推定のためのモデルの構築も課題となる．先に述べたように，寸法指標にはピットマンモデルの当てはまりの良いことが知られているが，多重ピットマンモデルをここで扱っているような多重寸法指標へ当てはめるのは難しい．モデルを用いるメリットは推定の安定の他に計算時間が短い点も挙げられる．大きな母集団を扱う場合を考えるとモデルの構築が必要と考えられる．

謝辞

本稿は，日本学術振興会科学研究費補助金（16H02791，18H00835），統計数理研究所共同利用研究（30-共研-2031，2019-ISMCRP-2042，2019-ISMCRP-4305）の研究成果に基づくものである．

参考文献

[1] 伊藤伸介（2010）．ミクロデータにおける秘匿性の評価方法に関する一考察，明海大学経済学論集，**22**（2），1-17.
[2] 伊藤伸介，村田磨理子，高野正博（2014）．ミクロデータにおける匿名化技法の適用の可能性－全国消費実態調査と家計調査を用いて－，統計研究彙報，**71**，83-124.
[3] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions, *Probability Theory and Related Fields*, **102**, 145-148.
[4] 佐井至道（2003）．母集団寸法指標のノンパラメトリック推定，統計数理，**51**（2），183-197.
[5] 佐井至道（2008）．母集団多重寸法指標のノンパラメトリック推定，岡山商大論叢，**43**（3），1-18.
[6] 佐井至道（2009）．多重母集団寸法指標のノンパラメトリック最尤推定－2 時点の個票データへの適用－，統計数理，**57**（2），425-442.
[7] 佐井至道（2013）．官庁統計におけるセンシティブデータ，システム／制御／情報，**57**（4），147-152.
[8] 佐井至道（2015a）．ノイズが加えられた個票データのリスク評価，2015 年度統計関連学会連合大会予稿集，87.

[9] 佐井至道（2015b）．ノイズが加えられた個票データのリスク評価，研究集会「官庁統計データの公開における諸問題の研究と他分野への応用」報告資料，1–5.
[10] 佐井至道（2016）．ノイズが挿入された個票データのリスク評価，岡山商大論叢，**52**（1），23–50.
[11] 佐井至道（2017a）．曖昧な母集団情報を考慮に入れたノイズを含む個票データのリスク評価，岡山商大論叢，**53**（1），27–57.
[12] 佐井至道（2017b）．ノイズが挿入された個票データの変数の型によるリスクの差について，2017年度統計関連学会連合大会予稿集，81.
[13] 佐井至道（2017c）．個票データに挿入するノイズの型によるリスクの差について，研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」報告資料，1–4.
[14] 佐井至道（2018a）．個票データのキー変数の型と開示リスクとの関係，岡山商大論叢，**54**（1），1–28.
[15] 佐井至道（2018b）．リスクの高いデータの秘匿について，2018 年度統計関連学会連合大会予稿集，18.
[16] 佐井至道（2018c）．秘匿措置が施されたデータのリスク評価の試み，研究集会「公的大規模データの利用におけるプライバシー保護の理論と応用」報告資料，1–5.
[17] 渋谷政昭（2005）．滑らかな罰金関数，科学研究費補助金研究会資料，1–3.
[18] 渋谷政昭，佐井至道（2007）．多重指標の確率分割，応用統計学会2007年度年会講演予稿集，13–18.
[19] Sibuya, M. and Sai, S. (2008). Analysis of a Dataset for Statistical Disclosure Control by Random Partition of a Multi-index, *Cherry Bud Worlshop 2008*, 1–13.
[20] U. S. Census Bureau (2014). Census 2010 Public Use Microdata Sample (microdata), Washington, D. C.
[21] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*, Springer.
[22] Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, Springer.